基于语义相似度的文本聚类研究*

毕　强[1]　刘　健[1]　鲍玉来[1,2]
[1](吉林大学管理学院　长春 130022)
[2](内蒙古大学图书馆　呼和浩特 010021)

摘要:【目的】为解决传统的文本聚类无法充分挖掘文本资源语义信息以及相似度矩阵高维性、稀疏性等问题,并进一步改善文本聚类质量,提出基于语义相似度的文本聚类方法。【方法】通过《同义词词林扩展版》计算词语的语义相似度并得到文本语义相似度矩阵,根据文本语义相似度矩阵进行谱聚类,将文本聚集为文本簇。【结果】利用复旦大学文本语料库与搜狗文本语料库中的文本资源作为数据来源分别对传统聚类算法与本文提出的算法进行实验,结果表明,当聚类个数为10时,本文算法的准确率最高,并且Purity值高于传统聚类算法的Purity值。【局限】《同义词词林扩展版》中包含的领域术语不完整,部分相似度计算结果需要手工进行调整。【结论】该方法考虑了词语间语义关系,充分挖掘文本主体潜在信息,并且改善了聚类质量,为文本聚类和推荐提供了一条新途径。

关键词: 同义词词林扩展版　语义相似度　谱聚类　文本挖掘
分类号: G250.7

1　引　言

Web2.0 时代,文本数据呈现爆炸式增长[1]。文本聚类作为一种无监督的机器学习方法,可以对文本信息进行有效的组织、分类和导航[2],从而保证用户对知识进行有效、便捷的获取。然而,文本聚类过程中,采用向量空间模型计算文本间相似度的方法受共现特征词影响较大[3],易造成描述概念信号弱、噪音数据多及特征矩阵稀疏等问题[4];基于领域本体计算概念相似度的方法需要人工或半人工构建本体,构建过程复杂,借助领域专家和知识工作人员协作完成,并且本体结构中包含信息较为复杂,不能充分体现和揭示概念之间的语义关系,相似度计算结果精度不高[5]。另外,在文本聚类中也存在着对初始聚类中心选值的敏感性问题、容易陷入局部最优值等问题[6],影响了文本聚类效果。

《同义词词林扩展版》编码简单,层次结构清晰,具有丰富的语义知识并且可以解决中文文本多义词分歧的问题[7],因此本文利用同义词词林扩展的语义相似度计算方法改进谱聚类算法:通过同义词词林计算语义相似度并形成语义相似度矩阵,对语义相似度矩阵进行拉普拉斯变换以降低矩阵维度,将变换后的向量矩阵进行聚类,从而完成对语义相近文本簇的划分,以此提高文本聚类效果。

2　相关研究

2.1　语义相似度计算

概念语义相似度是指两个概念间的相似程度[8],已经被应用于词义消歧[9]、自动检索[10]、图像分类及标注[11]、信息抽取[12]、信息检索[13]等领域。目前,语义相似度计算方法主要包括基于本体的概念语义相似度计算与基于语义词典的概念相似度计算。基于本体的语义相似度计算按照计算方法的不同可分为:基于距离的方法、基于内容的方法和基于属性的方法等。基于距离的计算方法是在层次网络中使用路径长度来

通讯作者:鲍玉来, ORCID:0000-0003-2528-5412, E-mail:65003846@qq.com。
*本文系国家自然科学基金项目"语义网络环境下数字图书馆资源多维度聚合与可视化展示研究"(项目编号:71273111)的研究成果之一。

量化两个概念之间的语义距离[14]。基于属性的方法[15]是利用事物之间不同的属性特征区别事物。两个事物的公共属性越多，相似度越高。基于内容的方法[16]认为两个概念共享的信息会影响二者的语义相似度。然而由于本体结构中包含信息较为复杂，不能充分体现和揭示概念之间的语义关系，导致相似度计算的精度不高。另一方面，利用语义词典 WordNet FrameNet、MindNet 等来计算英文词语相似度，以及利用《知网》(HowNet)、同义词词林等计算中文相似度[17]，也是较为常用的方法。基于语义词典的方法通常依赖于比较完备的大型语义词典。词典中的关系和层次结构，如概念之间的上下位关系和同位关系可以用来计算词语的相似度。由于基于同义词词林比基于《知网》的词汇语义相似度计算方法更符合人们的理解[18]，因此本文利用其作为计算语义相似度的方法。

2.2 文本聚类分析

文本聚类分析是利用文本之间的相似性对无结构或半结构化的文本对象进行自动分组的过程[19]。同组内文本相似性较高，不同组的文本相似性较低。通常将文本表示成向量的模式，利用特征词来计算各文本之间的相似度。常用的文本聚类分析的方法包括 K-means 聚类[20]、层次聚类[21]、基于密度的聚类[22]以及基于网格的聚类[23]等。文本聚类的过程包括提取文本特征词、计算文本相似度以及文本聚类算法等方面。文本聚类技术在文档整理、组织以及信息检索中得到广泛应用，例如对网页自动归类、新闻报道自动分类、电子邮件分组等，还可以对搜索引擎返回的结果进行聚类，使用户迅速查询到所需要的信息。

传统的聚类算法都是建立在凸球形的样本空间上。当样本空间不为凸时，算法会陷入“局部”最优。另外，许多文档之间没有公共词语存在，导致文档矩阵具有高维性和稀疏性，而且聚簇中心也没有提供可以理解的聚簇描述。为了能在任意形状的样本空间上聚类，收敛于全局最优解，克服文档矩阵的高维性和稀疏性等缺点，相关学者开始利用谱方法来聚类。谱聚类方法建立在谱图理论的基础上，通过计算数据相似关系建立相似度矩阵，以该矩阵的前 k 个特征向量来对不同的数据点聚类。与其他聚类方法不同，谱聚类不容易陷入局部最优解，而且可以有效识别非凸分布的聚类，已经成功应用于在线学习分类[24]、图像分割[25]、词义消歧[26]、网页划分[27]和文本挖掘[28]等领域。因此，本文选用谱聚类作为文档聚类的分析方法。

3 计算方法及过程

文本聚类过程中首先要对文本文档数据进行预处理，完成从文本形式到数学表示的转换。常见的文本表示方法采用向量空间模型，利用单词或词语共现次数表征文档内容，忽略了文档资源之间存在的语义关联。基于距离的相异度可以用来度量文档对象之间相似度，例如余弦距离、欧几里德距离、曼哈坦距离等。但由于文档之间的特征词交集过少导致文档向量矩阵的高维性和稀疏性，距离度量往往不能准确有效地表达文档之间潜在的语义关联信息。因此，在文本聚类过程中应充分挖掘隐藏在文档中的语义信息，寻找文本对象之间特有的语义关联。本文利用改进的语义相似度矩阵代替空间向量模型，并利用谱聚类方法对相似度矩阵进行分解，从而降低矩阵的高维度，提高聚类结果的准确性。

3.1 概念语义相似度计算

文献[18]根据同义词词林结构及其编排的特点，利用词语在词林树状结构中的编号，提出基于同义词词林的概念语义相似度计算方法。本文参考文献[18]的计算方法计算概念的语义相似度。具体描述如下：首先判断两个概念在同义词林中不同编号的起始位置，例如：Aa01A01 与 Aa01B01，在第四层不同。对于不同的层，分别乘以不同的系数。同义词词林的结构深度共五层，从第二层开始，对于不同层的词语分别乘以不同的参数 a、b、c、d。然后再乘以调节参数 $\cos\left(n\times\frac{\pi}{180}\right)$，利用该调节参数将词语相似度控制在[0, 1]区间，其中 n 是分支层的节点总数。

概念所在词林位置的密度会影响概念语义的相似度计算：密度越大，概念语义相似度的值越精确；相反，密度越小，概念语义相似度值误差越大。一般的计算方式是统计两个概念在词典间隔单词的数量，即计算词林中公共祖先的数量来计算概念语义相似度，这种方法并没有考虑概念所在分支的密度信息。通过统计两个概念 c_1, c_2 在同义词词林中分支间的距离，即统计这两个概念所在分支包含的概念数量来计算密度信息[29]，密度信息公式如下。

$$\text{dis} = -\log\left(\frac{\text{freq}(c)}{N}\right) \quad (1)$$

其中，freq(c)=$\sum$count(c)，c 为从概念 c_1 所在分支到概念 c_2 所在分支之间所包含的概念，$\sum$count(c)为这些概念数量的总合，N 为 c_1 和 c_2 所在分支的所有概念的总和。利用公式(1)对计算的语义相似度结果进行细化，以此保证计算结果更加精确。由以上得出概念的语义相似度公式，用 Sim 表示。

若两个概念不在同一棵树上：

$$\text{Sim}(c_1,c_2) = f \quad (2)$$

若两个概念在同一棵树上，并且位于在第二层分支，则系数为 a，计算公式如下：

$$\text{Sim}(c_1,c_2) = 1\times a\times\cos\left(n\times\frac{\pi}{180}\right)\times \text{dis} \quad (3)$$

若两个概念在同一棵树上，并且位于第三层分支，则系数为 b，计算公式如下：

$$\text{Sim}(c_1,c_2) = 1\times b\times\cos\left(n\times\frac{\pi}{180}\right)\times \text{dis} \quad (4)$$

若两个概念在同一棵树上，并且位于第四层分支，则系数为 c，计算公式如下：

$$\text{Sim}(c_1,c_2) = 1\times c\times\cos\left(n\times\frac{\pi}{180}\right)\times \text{dis} \quad (5)$$

若两个概念在同一棵树上，并且位于第五层分支，则系数为 d，计算公式如下：

$$\text{Sim}(c_1,c_2) = 1\times d\times\cos\left(n\times\frac{\pi}{180}\right)\times \text{dis} \quad (6)$$

当编号相同且末尾号为“=”时，相似度为 1；当编号相同而只有末尾号为“#”时，直接将定义的系数 e 赋给结果。即：$\text{Sim}(c_1,c_2)=e$。通过对概念相似度测试及根据文献[18]的参考，本文将层数初始值设置为 a = 0.532, b = 0.78, c = 0.84, d = 0.88, e = 0.42, f = 0.001。

3.2 文本相似度计算

文本相似度是指文本间主题或内容的相似程度，与 Quillian 的联合概念相似，可以通过计算文本特征词或概念的相似度计算文本相似度[30]。当计算文本的语义相似度时，首先要计算文本的语义距离，如公式(7)[30]所示。

$$\text{Dist}(d_x,d_y) = \text{Dist}(\wedge_{i=1}^{n}K_{xi},\wedge_{j=1}^{m}K_{yj}) = \frac{1}{d}\sum_{i=1}^{n}\sum_{j=1}^{m} f_i\times f_j\times \text{Dist}(K_i,K_j) \quad (7)$$

其中，d_x，d_y 为两个不同文本，x_i，y_j 分别为文本 d_x，d_y 所包含的特征词或概念；f_i 为概念 x_i 在文本 dx 中出现的次数；f_j 为概念 y_j 在文本 d_y 中出现的次数；n，m 分别为两个文本所包含的概念个数。为了避免语义距离的计算结果过大，利用 d 进行归一化，公式如下[30]。

$$d = \left(\sum_{i=1}^{n} f_{xi}\right)\times\left(\sum_{j=1}^{m} f_{yj}\right) \quad (8)$$

d 所代表的意义为两个文本中概念或者特征词语义距离的数量，同时也考虑到特征词或概念在文本中出现的次数，对语义距离进行归一化可以避免文本包含的特征词或概念过多，导致文本语义距离过大。综上，本文将文本语义相似度定义如下[30]。

$$\text{Sim}(d_x,d_y) = \frac{1}{1+\text{Dist}(d_x,d_y)} \quad (9)$$

可以看出，语义距离越大，文本的相似度越小。

3.3 基于语义相似度的谱聚类算法

基于 NJW 算法[31]，本文提出基于语义相似度矩阵的谱聚类算法(SCBSS)。SCBSS 算法采用概念列表表示文本，以文本间的语义相似度作为文本间相关程度的度量。相似度矩阵是一个对称矩阵，而且相似度值是非零的。在进行文本预处理的基础上，以中文词语为单位，利用《同义词词林扩展版》计算词语之间的语义相似度，将其作为衡量概念距离的指标。其次，将文本表示成概念的集合，两个文本的相似度可以通过它们包含概念的语义相似度计算。最后，构建文本间相似度矩阵，并应用文本谱聚类方法进行分析。改进 SCBSS 算法的描述如下：

输入：n 个数据点，聚类的个数 K
输出：K 个聚类
方法：
Begin
①构造相似性矩阵 $W\in R^{n\times n}$；
②构造矩阵 $P = D^{-1/2}WD^{1/2}$；
③求 P 的 k 个最大特征值所对应的特征向量 $v_1, v_2, \cdots, v_n$，构造矩阵 $V=[v_1v_2\cdots v_n]\in R^{n\times k}$，其中 v_i 为列向量，i=1, …, n;
④规范化 V 的行向量，得到矩阵 Y，其中 $y_{ij} = v_{ij}/(\sum_j v_{ij}^2)^{1/2}$；
⑤将 Y 的每一行看成是 R^k 空间中的一点，使用 K-means 聚类。
End

如上所示，谱聚类将文档的相似度放到一个带权无向图中，采用“图划分”的方法进行聚类。谱聚类算

法分为三步:

(1) 构造一个 $n \times n$ 的权值矩阵 W, 利用同义词词林计算词语的相似度 w_{ij}。w_{ij} 表示词语 i 和词语 j 的相似度, 显然 W 是对称矩阵。

(2) 构造一个对角矩阵 D, d_{ii} 为 W 第 i 列元素之和。对矩阵 P 进行规范化, 即 $P = D^{-1/2}WD^{1/2}$。可以证明 P 是个半正定和对称矩阵, 求得 P 的前 n 大特征值所对应的特征向量。

(3) 将 n 个特征向量放在一起构造一个 $n \times k$ 的矩阵 V, 将 V 的每一行当成一个新的样本点, 对新的样本点进行 K-means 聚类。传统的聚类方法要求数据必须是 N 维欧氏空间中的向量, 而利用谱方法聚类只需要计算文本的相似度矩阵, 这降低了数据矩阵的维度, 并且缓解了数据矩阵的稀疏性。

4 实验过程及结果

4.1 语义相似度计算

选取 10 组概念进行语义相似度计算, 为了对比实验效果, 采用咨询的方式获得人工对于语义相似度的判断。咨询对象包括计算机专业、情报专业、经济专业的硕士生和博士生, 共有 20 人。通过对该组概念语义评价问题进行语义相似度判定。语义相似度的评判范围是[0, 1], 0 表示两个概念完全不同, 1 表示两个概念语义相同。对受测者各进行两次实验, 并对同一概念语义相似度的评测结果取平均值。计算结果如表 1 所示。

表 1 概念语义相似度部分计算结果

词语		语义相似度	相似度范围
经济	产业	0.693	0.7-0.8
货币	银行	0.540	0.6-0.7
企业	公司	1.000	0.9-1.0
资源	工业	0.362	0.3-0.4
软件	计算机	0.717	0.7-0.8
服务器	路由器	0.500	0.5-0.6
历史	二战	0.431	0.4-0.5
地理	地理学	1.000	0.9-1.0
电路	电子	0.832	0.8-0.9
设备	产品	0.727	0.7-0.8

该方法根据概念在同义词词林的位置进行编码, 计算得出概念相似度。从表 1 可以看出, 利用同义词词林进行语义相似度计算结果具有较高的准确性, 并且符合目标用户对于语义相似度的主观判断, 说明该算法可以客观准确地反映概念之间的语义关系, 并为有效度量概念的语义相似度提供一种新的方法和途径。

4.2 文本相似度计算结果

搜狗文本挖掘数据集是比较全面的语料库, 该数据集包含汽车、财经、IT、健康、体育、旅游、教育、招聘、文化、军事等 10 个类别, 每个类别大约有 2 000 篇文档。本文从这 10 个类别中各选择 100 篇文档共计 1 000 篇, 利用 NLPIR 大数据搜索与挖掘共享平台① 对其进行分词处理和词频统计。从中选出 10 个词频较高并能代表文档内容的关键词, 将其作为表征文档特征的关键词, 并记录其词频, 如图 1 所示。

文章名	关键词	词频	关键词	词频	关键词	词频	关键词	词频	关键词	词频	关键词	词频	关键词	词频	关键词	词频	关键词	词频	关键词	词频	关键词	词频
文章1	流通	166	经济学	101	理论	72	商业	31	经济	42	资源	8	马克思	23	国家	7	变革	6	政府	5	要素	12
文章2	企业	77	经济	25	商业	33	竞争性	10	市场	21	结构	10	行业	27	资产	10	机制	9	产业	5	资本	4
文章3	价格	7	市场	18	消费者	10	企业	8	购买力	3	居民	10	制度	6	体制	3	资金	3	指数	5	医疗	4
文章4	农村	86	农民	79	农产品	62	消费	47	市场	40	收入	40	城镇	37	支出	20	价格	18	投资	13	资源	4
文章5	消费	56	储蓄	42	居民	29	经济	20	存款	17	投资	10	市场	7	改革	6	银行	6	消费品	4	余额	4
文章6	消费	118	经济	58	增长	41	投资	37	储蓄	27	需求	17	收入	17	政策	13	财政	7	产业	7	信贷	5
文章7	消费	72	发展	36	经济	29	生产力	17	生产关系	7	社会	5	消费者	5	福利	5	政策	4	社会	5	管理	4
文章8	经济	28	市场	23	消费	23	商家	14	行业	10	服务	9	发展	7	产业	6	需求	6	政府	6	内需	3
文章9	企业	72	信息	65	网络	63	外贸	30	经济	25	管理	22	市场	19	资源	12	成本	9	技术	9	服务	8
文章10	俄罗斯	53	经济	25	出口	23	经贸	18	经济危机	13	政策	13	市场	11	贸易	10	财政	6	工业	6	投资	6
文章11	投资	35	财政	28	经济	18	需求	17	贴息	12	企业	11	财政	8	消费	7	投资	14	产业	6	债券	5
文章12	储蓄	225	政府	194	支出	79	收入	47	税收	47	投资	32	经济	30	资金	25	政策	24	财政	22	财政赤字	16
文章13	财政	88	政策	55	国债	30	经济	23	预算	22	政府	20	投资	19	货币	16	银行	16	企业	12	市场	9
文章14	财政	56	经济	26	支出	19	预算	15	资金	15	社会	11	收入	8	投资	8	企业	6	国民经济	4	政治	5
文章15	税收	9	征管	5	制度	5	电子商务	3	交易	3	经济	3	经济体制	3	市场	3	保障	2	改革	2	贸易	2
文章16	资本市场	42	政策	24	货币	21	经济	17	市场	15	投资	14	金融	13	产业	11	企业	9	股市	6	银行	6
文章17	税收	29	协定	12	经济	8	技术	7	金融	7	资本	6	税务	5	电子商务	4	纳税人	4	劳动力	3	税基	3
文章18	金融	64	经济	28	知识	19	货币	16	市场	12	金融业	7	产业	6	企业	5	资本市场	5	经济体制	4	商业	4
文章19	银行	73	金融	15	经营	11	利润	8	企业	7	服务	6	农业	5	贷款	4	金融业	4	商业	4	货币	3
文章20	银行	38	金融	30	信息化	14	科技	12	经济	9	银行业	7	计算机	4	信息	4	货币	4	电子商务	3	外资	3
文章21	金融	273	银行	77	贷款	60	资产	43	企业	30	经济	26	商业	21	制度	18	资金	13	金融市场	12	金融业	8
文章22	保险	93	失业	88	企业	45	基金	22	经济	11	创新	10	改革	10	服务	9	财政	5	市场经济	5	公有制	4
文章23	企业	78	技术	53	经济	34	产业	28	知识经济	28	信息	28	信息化	25	市场	17	公司	13	资本	13	工业	8
文章24	经济	67	法律	29	经济法	24	立法	17	社会	17	利益	12	行政	9	私法	7	改革	6	产品	5	服务	5
文章25	经济	105	社会主义	45	生产力	18	现代化	16	政治	14	改革	13	核心	13	阶级斗争	12	改革开放	7	国有经济	7	国民经济	5
文章26	利益	73	集体主义	47	社会主义	46	市场经济	39	个人主义	35	道德	29	观念	12	经济	12	资产阶级	12	政治	9	历史	5
文章27	经济	85	市场	57	市场经济	55	企业	54	伦理	53	社会主义	45	道德	42	自由	29	竞争	25	权利	19	民主	17
文章28	经济	59	知识	41	生产力	33	社会	25	资源	16	工业	14	产业	11	农业	9	资本	8	技术	7	劳动	7
文章29	道德	109	建设	31	观念	21	社会主义	20	市场经济	13	政策	12	经济	6	思想	5	市场	4	需求	4	优势	2
文章30	政治	31	思想	22	军队	16	利益	10	社会主义	8	教育	6	素质	6	制度	5	部队	4	国家	4	使命	4

图 1 文档分词及词频统计结果

①http://ictclas.nlpir.org/nlpir/.

利用 Java 语言对公式(7)–公式(9)进行编程计算语义相似度及文本相似度。部分计算结果如图 2 所示。

	文章1	文章2	文章3	文章4	文章5	文章6	文章7	文章8
文章1	1.000000000000000	0.563974568901669	0.327679511725551	0.94941099423074	0.1005755767596925	0.9728397634991666	0.969634158229821	0.666262179435328
文章2	0.563974568901669	1.000000000000000	0.766954062882087	0.675106095729467 4	0.163081826396616	0.4019696407249738	0.829608338463549	0.165699911380247
文章3	0.3276795117255515	0.766954062882087	1.000000000000000	0.02636901259048 1	0.90425783123187	0.918461274220261	0.175794923419562	0.6359152355969346
文章4	0.9494109942307407	0.675106095729467	0.026369012590482	1.000000000000000	0.420746187724109	0.6387285229573507	0.017402323041681	0.4130173790115754
文章5	0.100575576759693	0.163081826396616	0.904257831231872	0.4207461877241 1	1.000000000000000	0.594445989739732	0.594000170522666	0.8515256634203805
文章6	0.972839763499167	0.401969640724974	0.9184612742202605	0.6387285229573503	0.594445989739732	1.000000000000000	0.71068311265961	0.428263053220396
文章7	0.9696341582298214	0.82960833846355	0.175794923419562	0.017402323041681	0.594000170522666	0.71068311265961	1.000000000000000	0.4354999400392616
文章8	0.6662621794353285	0.1656999113802464	0.635915235596934	0.413017379011575	0.8515256634203796	0.4282630532203955	0.435499940039262	1.000000000000000
文章9	0.055181703157101	0.856106829731903	0.1422756519163	0.033623373117416	0.470549414597219	0.8835696677065017	0.132718341869498	0.416412334885247
文章10	0.310603435121219	0.2612428331351793	0.369832290345109	0.2367970011278646	0.7924533845516057	0.971914982838976	0.366547439196948	0.7936447365840604
文章11	0.438117547541325	0.823865464416473	0.519229085912907	0.104859881506144	0.102005203451274	0.124725826624569	0.263593179253916	0.6897209873570302
文章12	0.279399821825014	0.6057667862248026	0.192426142122578	0.111364329026387	0.8047438312755535	0.8017474051834266	0.998475996449874	0.666745668415216
文章13	0.49028001555837	0.672217918962981	0.457340389979846	0.082153321463477	0.956683166986077	0.9685143918589274	0.077561997403399	0.684369100937615
文章14	0.37548598939755	0.698221902673603	0.243407145830699	0.177327074492063	0.984791505749781	0.8235229782207427	0.02093211409402	0.538429440587028
文章15	0.361509827348179	0.328644826509014	0.134085763839688	0.587833755428259	0.527434572676645	0.847792529417909	0.813479130677271	0.138275549760936
文章16	0.985339344546746	0.994287926160746	0.84254223138336	0.351662037701017	0.3190870346828705	0.425320867189221	0.426701550674164	0.0023055906289375
文章17	0.044912769652934	0.7949372982548635	0.114587787531415	0.732860607433933	0.5411675933829345	0.8219995783201957	0.981793927895194	0.547719869515687
文章18	0.060415667541866	0.115359529233375	0.957269607765617	0.546111166215435	0.521022474081615	0.638057711978314	0.613397631820954	0.050689166371367
文章19	0.147779629408661	0.398509443532778	0.017006164162632	0.575206097071224	0.5929222575998 5	0.6286251667642055	0.7627835000836285	0.0505777926001265
文章20	0.7783298361351823	0.345390713589556	0.651282051403014	0.604412949194359	0.125158518524769	0.4233416049016587	0.590354267691677	0.26378988255054
文章22	0.402738381718552	0.138404523698873	0.445280382237652	0.143303767417171	0.080275315952335	0.1715648900044044	0.3111443475433875	0.8032132851294214
文章22	0.10483724127857	0.5780178080372598	0.080193262837647	0.6894201471918455	0.743499629038064	0.9008811990118084	0.8338514033022233	0.3453715985749812
文章23	0.7110146136680027	0.393788514277369	0.4973729245975607	0.412359724815352	0.642414148314466	0.260937182014793	0.6809330690274096	0.7862874616627886
文章24	0.852558624950651	0.868882452292448	0.784765437278941	0.342651302252070 5	0.472137768645501	0.406217655488023	0.348471586841287	0.932447040678290 8
文章25	0.4838700311251856	0.9023832838666332	0.345528483590705	0.110038349717243	0.671893442927284	0.1206465875359752	0.1543906140607096	0.9716173605719427
文章26	0.8385271683015687	0.445333670840423	0.7883641667853447	0.601505462390621 7	0.269920483401144	0.451679716214165	0.7197404083241876	0.270232890868706
文章27	0.056551202951806	0.393540011663058	0.008005874011931	0.585764152008852	0.713331270454713	0.698449684405485	0.7612922589034135	0.9673682921536115
文章28	0.448030401984392	0.002832992183949	0.3736750531365427	0.1367883801642233	0.674255808423912	0.1896826343631757	0.254612478836745	0.9633271849773037
文章29	0.602312025010499	0.207572351849996	0.639533911828526	0.151952532920807	0.389387454231344	0.253845526997306	0.390054097939726	0.9379319034155484
文章30	0.010089965720908	0.255517342110042	0.8670937216539683	0.590271889861161	0.549722995873713	0.771690193392097	0.576826431770618	0.318361222785512
文章31	0.253213354239287	0.405561566226373	0.3722728725263376	0.359119957039426	0.727446294215676	0.026305175466306	0.5044707218883255	0.805644349331044
文章32	0.544157812908578	0.709817986907104	0.27898522136654	0.217660147899522	0.920681087249042	0.8541084609723018	0.096874772727638	0.5251280427283183
文章33	0.410493040486498	0.604007500171299	0.352302557400517	0.9363983206959867	0.949085761487553	0.025088237336723	0.9823815271374827	0.5589581019354415

图 2　文档相似度计算结果

4.3　文本聚类实验及结果

得到文本相似度矩阵后，利用谱聚类方法对文档矩阵进行聚类划分。本文聚类结果的衡量指标选用聚类结果的纯度(Purity)进行分析，此方法是一种简单有效的聚类结果的评价指标，计算公式如下。

$$Purity = \sum_{i=1}^{K} \frac{m_i}{m} p_i \quad (10)$$

其中，$p_i = \max(p_{ij})$，$p_{ij} = \frac{m_{ij}}{m_i}$，$m_i$ 是在聚类 i 中所有成员的个数，m_{ij} 是既属于聚类 i 又属于聚类 j 的成员个数。对于该算法，分别考虑当聚类个数 K=4、10 这两种情况，对每个取值均随机选择初始簇中心，得到聚类结果。当 K=4 时，聚类的结果如表 2 所示。

表 2　K=4 时聚类结果

类别	C1	C2	C3	C4
汽车	82	4	5	9
财经	10	64	12	14
IT	40	17	24	19
健康	19	17	33	31
体育	37	15	10	38
旅游	23	35	18	24
教育	8	3	84	5
招聘	46	39	9	7
文化	2	13	8	77
军事	28	24	30	28

由表 2 所示，通过算法聚类后，得到 Purity=0.307。当 K=10 时，聚类的结果如表 3 所示。

表 3　K=10 时聚类结果

类别	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
汽车	67	8	4	3	0	2	2	6	1	7
财经	2	73	5	3	3	4	1	2	5	2
IT	5	4	69	2	3	8	1	3	2	3
健康	7	4	3	57	10	4	2	5	3	5
体育	0	2	2	0	92	1	1	0	2	0
旅游	1	4	2	1	6	78	3	1	2	2
教育	2	6	1	9	2	1	62	5	6	4
招聘	3	3	3	1	2	1	1	84	1	1
文化	2	4	5	5	8	7	11	4	55	9
军事	1	2	1	1	2	2	1	1	1	88

可以看出，聚类个数 K=10 时 Purity= 0.725。这说明随着聚类个数的增多，聚类的结果越来越准确。

选取 K-means、TCUSS[16]以及本文提出的 SCBSS 三种方法进行实验，并将聚类个数设置为 4、5、6、7、8、10，对于不同的聚类数量各重复实验 10 次，然后取 Purity 值的平均数集 $C(P)=\sum f/10$，计算结果如表 4 所示。

表 4　三种聚类算法的 Purity 值对比

聚类数量	K-means	TCUSS	SCBSS
4	0.296	0.303	0.307
5	0.272	0.411	0.439
6	0.371	0.506	0.565
7	0.433	0.517	0.688
8	0.466	0.513	0.706
10	0.483	0.504	0.725

如表4所示，SCBSS算法为本文的算法，TCUSS、K-means为人工方式构建。结果表明，SCBSS算法的纯度有明显提高。并且当聚类数量较少时，算法的Purity值并不高，但是随着聚类数量的增多，Purity值有显著提升。由于SCBSS算法采用概念列表示文本，并基于《同义词词林》语义相似度计算方法对文本进行相似度计算，解决了基于向量空间模型的文本聚类算法中数据维数过高和相似度矩阵稀疏等问题，也解决了文本中包含的近义词和多义词问题，从而提高了聚类的效果和质量。但是，在对表征本文关键词的选取过程中由于个人主观因素的差异导致关键词选取不准确，聚类结果的精确度计算出现偏差；另外，《同义词词林扩展版》作为一种语义资源，存在未登录词的问题，互联网语料库中很多新词需要人工标示其相似度，由此也会影响聚类结果。以上问题也是今后研究的重点。

5 结 语

本文提出基于语义相似度的文本聚类方法SCBSS。首先，对文本进行预处理，提取出文本的特征词，利用《同义词词林扩展版》进行词语间的语义相似度计算，以此作为计算文本间相似度的依据，并构造文本相似度矩阵。其次，对相似度矩阵进行规范化，求得最大特征值以及对应的特征向量，并构造特征向量矩阵。最后，使用谱聚类方法对新的特征向量构成的矩阵进行聚类，完成文本的划分。相对于基于本体的方法计算语义相似度，本文提出的基于《同义词词林扩展版》的语义相似度计算方法的计算结果更加准确。利用本文提出的谱聚类方法，解决了传统聚类算法数据维数过高和矩阵稀疏等问题。实验结果表明，SCBSS可以充分挖掘聚类中文本之间的语义相似度，同时提高了聚类结果的质量。

参考文献：

[1] 王鹏，高铖，陈晓美．基于LDA模型的文本聚类研究[J]．情报科学，2015，33(1)：63-68．(Wang Peng, Gao Cheng, Chen Xiaomei. Research on LDA Model Based on Text Clustering[J]. Information Science, 2015, 33(1): 63-68.)

[2] 顾晓雪，章成志．结合内容和标签的Web文本聚类研究[J]．现代图书情报技术，2014(11)：45-52．(Gu Xiaoxue, Zhang Chengzhi.Using Content and Tags for Web Text Clustering [J]. New Technology of Library and Information Service, 2014(11): 45-52.)

[3] 赵辉，刘怀亮．面向用户生成内容的短文本聚类算法研究[J]．现代图书情报技术，2013(9)：88-92．(Zhao Hui, Liu Huailiang.Research on Short Text Clustering Algorithm for User Generated Content [J]. New Technology of Library and Information Service, 2013(9): 88-92.)

[4] 柴春梅．互联网短文本信息分类关键技术研究[D]．上海：上海交通大学，2009．(Chai Chunmei. The Key Technology Research on Internet Short Text Information Classification [D]. Shanghai: Shanghai Jiaotong University, 2009.)

[5] 张文秀，朱庆华．领域本体的构建方法研究[J]．图书与情报，2011(1)：16-19，40．(Zhang Wenxiu, Zhu Qinghua. Research on Construction Methods of Domain Ontology [J]. Library and Information, 2011(1): 16-19, 40.)

[6] 行小帅，潘进，焦李成．基于免疫规划的K-means聚类算法[J]．计算机学报，2003，26(5)：605-610．(Xing Xiaoshuai, Pan Jin, Jiao Licheng. A Novel K-means Clustering Based on the Immune Programming Algorithm [J]. Chinese Journal of Computers, 2003, 26(5): 605-610.)

[7] 刘端阳，王良芳.基于语义词典和词汇链的关键词提取算法[J]．浙江工业大学学报，2013，41(5)：545-551．(Liu Duanyang, Wang Liangfang. Keywords Extraction Algorithm Based on Semantic Dictionary and Lexical Chain [J]. Journal of Zhengjiang University of Technology, 2013, 41(5): 545-551.)

[8] 刘宏哲，须德．基于本体的语义相似度和相关度计算研究综述[J]．计算机科学，2012，39(2)：8-13．(Liu Hongzhe, Xu De. Ontology Based Semantic Similarity and Relatedness Measures Review [J]. Computer Science, 2012, 39(2): 8-13.)

[9] Fernandez-Amoros D, Heradio R.Understanding the Role of Conceptual Relations in Word Sense Disambiguation [J]. Expert Systems with Applications, 2011, 38(8): 9506-9516.

[10] Alonso I, Contreras D. Evaluation of Semantic Similarity Metrics Applied to the Automatic Retrieval of Medical Documents: An UMLS Approach [J]. Expert Systems with Applications, 2016, 44 (C): 386-399.

[11] Chang J Y, Lee K M. Large Margin Learning of Hierarchical Semantic Similarity for Image Classification [J]. Computer Vision and Image Understanding, 2015, 132: 3-11.

[12] Hassan H, Hassan A, Emam O. Unsupervised Information Extraction Approach Using Graph Mutual Reinforcement [C]. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.2006: 501-508.

[13] Bae M, Kang S, Oh S. Semantic Similarity Method for Keyword Query System on RDF [J]. Neurocomputing, 2014,

146(C): 264-275.

[14] Rada R, Mili H, Bicknell E, et al. Development and Application of a Metric on Semantic Nets [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1989, 19(1): 17-30.

[15] Tversky A. Feature of Similarity [J]. Psychological Review, 1977, 84(4): 327-352.

[16] Lord P W, Stevens R D, Brass A, et al. Investigating Semantic Similarity Measures Across the Gene Ontology: The Relationship Between Sequence and Annotation [J]. Bioinformatics, 2003, 19(10): 1275-1283.

[17] 焦芬芬. 基于概念和语义相似度的文本聚类算法[J]. 计算机工程与应用, 2012, 48(18): 136-141. (Jiao Fenfen. Clustering Method Based on Concept and Semantic Similarity [J]. Computer Engineering and Applications, 2012, 48(18): 136-141.)

[18] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报: 信息科学版, 2010, 28(6): 602-608. (Tian Jiule, Zhao Wei. Words Similarity Algorithm Based on Tongyici Cilin in Semantic Web Adaptive Learning System [J]. Journal of Jilin University: Information Science Edition, 2010, 28(6): 602-608.)

[19] 王刚, 邱玉辉. 基于本体及相似度的文本聚类研究[J]. 计算机应用研究, 2010, 27(7): 2494-2497. (Wang Gang, Qiu Yuhui. Study on Text Clustering Based on Ontology Similarity [J]. Application Research of Computers, 2010, 27(7): 2494-2497.)

[20] Xiong S, Ji D. Exploiting Flexible-constrained K-means Clustering with Word Embedding for Aspect-phrase Grouping [J]. Information Sciences, 2016, 367-368: 689-699.

[21] Zhuo Z, Zhang X, Niu W, et al. Improving Data Field Hierarchical Clustering Using Barnes-Hut Algorithm [J]. Pattern Recognition Letters, 2016, 80(1): 113-120.

[22] Kumar K M, Reddy A R M. A Fast DBSCAN Clustering Algorithm by Accelerating Neighbor Searching Using Groups Method [J]. Pattern Recognition, 2016, 58: 39-48.

[23] YıldırımA A, Özdoğan C. Parallel WaveCluster: A Linear Scaling Parallel Clustering Algorithm Implementation with Application to Very Large Datasets [J]. Journal of Parallel and Distributed Computing, 2011, 71(7): 955-962.

[24] Langone R, Agudelo O M, De Moor B, et al. Incremental Kernel Spectral Clustering for Online Learning of Non-stationary Data [J]. Neurocomputing, 2014, 139(2): 246-260.

[25] Yang Y, Wang Y, Xue X. A Novel Spectral Clustering Method with Superpixels for Image Segmentation [J]. International Journal for Light and Electron Optics, 2016, 127(1): 161-167.

[26] Chifu A-G, Hristea F, Mothe J, et al. Word Sense Discrimination in Information Retrieval: A Spectral Clustering-based Approach [J]. Information Processing & Management, 2015, 52(2): 16-31.

[27] Ng A Y, Zheng A X, Jordan M I. Stable Algorithms for Link Analysis [C]. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001: 258-266.

[28] Singh K, Shakya H K, Biswas B. Clustering of People in Social Network Based on Textual Similarity [J]. Perspectives in Science, 2016, 8: 570-573.

[29] 吕立辉, 梁维薇, 冉蜀阳. 基于词林的词语相似度的度量[J]. 现代计算机, 2013(1): 3-6, 9. (Lv Lihui, Liang Weiwei, Ran Shuyang. A Method for Measuring Word Similarity Based on Cilin [J]. Modern Computer, 2013(1): 3-6, 9.)

[30] 孙爽, 章勇. 一种基于语义相似度的文本聚类算法[J]. 南京航空航天大学学报, 2006, 38(6): 712-716. (Sun Shuang, Zhang Yong. Clustering Method Based on Semantic Similarity [J]. Journal of Nanjing University of Aeronautics & Astronautics, 2006, 38(6): 712-716.)

[31] Ng A Y, Jordan M L, Weiss Y. On Spectral Clustering: Analysis and an Algorithm[A]. // Advances in Neural Information Processing Systems[M]. Cambridge, MA: MIT Press, 2002.

作者贡献声明:

毕强: 提出研究命题及研究思路;

刘健: 论文撰写及最终版本修订, 数据处理及实证研究;

鲍玉来: 论文修改。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: tomosliu9999@126.com。

[1] 刘健, 毕强. Resoures.csv. 文章分词统计表.

[2] 刘健, 毕强. Similarity.csv. 文章相似度计算结果.

[3] 刘健, 毕强. Similarity Questionnaire.doc. 相似度调查表.

[4] 刘健, 毕强. Result.csv. 文章聚类结果统计表.

[5] 刘健, 毕强. PurityResult.csv. 文章聚类 Purity 值统计表.

收稿日期: 2016-09-12
收修改稿日期: 2016-10-24

A New Text Clustering Method Based on Semantic Similarity

Bi Qiang[1] Liu Jian[1] Bao Yulai[1, 2]
[1](School of Management, Jilin University, Changchun 130022, China)
[2] (Inner Mongolia University Library, Hohhot 010021, China)

Abstract: [**Objective**] This paper proposes an algorithm based on semantic similarity to extract more information from the textual resources. [**Methods**] First, we calculated the semantic similarity of words with the Extended Dictionary of Synonyms, and then created a semantic similarity matrix. Second, we clustered the texts based on the new semantic similarity matrix. [**Results**] The proposed algorithm was examined with text corpus from Fudan University and the search engine Sogou. Compared to the traditional methods, the proposed algorithm achieved the highest precision rates and purity values (cluster number=10). [**Limitations**] Some partial similarity calculation results were manually adjusted due to the incomplete coverage of the Tongyici Cilin Extened Edition. [**Conclusions**] The proposed algorithm could extract more latent information from the texts, which is an effective method to cluster and recommend textual documents.

Keywords: Tongyici Cilin Extended Edition Semantic similarity Spectrum clustering Text mining

Clarivate Analytics 发布 2016 年高被引研究人员

2016 年 11 月中旬, 曾经是 Thomson Reuters 的知识产权与科学业务的 Clarivate Analytics 公司发布了年度高引用研究人员列表。该列表是引用分析的结果, 给出了一些科学家名单, 这些科学家的研究在他们各自的研究领域在全球有着重大的影响。

本次引用分析根据 2004 年 1 月至 2014 年 12 月这 11 年期间的高被引文献, 选择了 21 个自然科学和社会科学领域的共 3 000 多名研究人员。由来自 Clarivate Analytics 的文献计量专家根据数据进行分析得出结果。该引用分析使用世界领先的基于网络的研究分析平台 InCites™ Essential Science Indicators℠, 基于科学绩效指标、来自 Web of Science™ 的学术论文发表数量和引用数据这些趋势数据。

Clarivate Analytics 出品的高被引研究人员数据是世界大学学术排名(http://www.shanghairanking.com/index.html)的关键组成部分, 是全球顶尖大学中历史最悠久且最有影响力的年度调查之一。德国马克斯普朗克学会科学和创新研究部门文献计量学和社会学家 Lutz Bornmann 认为, “在定量研究评估领域, 几乎没有另一个免费访问的数据库, 可以像 Clarivate Analytics 出品的高被引研究人员列表那样为研究人员带来如此高的声誉。”

Clarivate 负责人 Jessica Turner 表示: “我们的高引用研究人员名单在学术和科学界赢得了全球尊重, 我们感到很自豪。”

访问 http://hcr.stateofinnovation.thomsonreuters.com 可以查看 2016 年高被引研究人员名单。

(编译自: https://librarytechnology.org/news/pr.pl?id=22031)

(本刊讯)